Un théorème sur l'équivalence de la valeur H entre langues

Gabriel G. Bès

Un théorème sur l'équivalence de la valeur H entre langues

Gabriel G. Bès
Groupe de recherches sur la condensation de l'information en langue naturelle (CILN)
Université Blaise-Pascal, Clermont II

Condenser, Adosa, Clermont-Ferrand, février 1980, n° 1, p. 97-100

Résumé

Les codes de réduction alphabétique proposent un système de codification des lettres d'un texte en langue naturelle qui tend à réduire le nombre de symboles binaires nécessaires à la représentation du texte, tout en sauvegardant la restitution de celui-ci de manière univoque. Les codes étudiés sont fondés sur le principe de codifier non les symboles d'un alphabet unique, mais les symboles qui correspondent aux unités qui apparaissent dans les différents contextes et qui déterminent une pluralité d'alphabets. Cet article présente un théorème qui montre que si les règles permettant de passer d'un alphabet à l'autre définissent une relation bijective entre les suites traduites, le coût de la codification, mesuré en symboles binaires, ne variera pas si on réussit à proposer une codification des deux langages dont les valeurs en symboles binaires soient respectivement identiques à H et à H', leur entropie optimale.

Voir aussi

Gabriel G. Bès. « Structuration linguistique et mesure de l'information. » *Linguistique fonctionnelle. Débats et perspectives*, présentés par M. Mahmoudian. Presses universitaires de France, Paris, 1979, p. 129-141. https://hal.archives-ouvertes.fr/hal-01100168

Gabriel G. Bès. « Description phonologique et codification économique du langage ». *Cahiers du Centre Interdisciplinaire des Sciences du Langage*, 1980, n° 2, p. 117-123. https://hal.archives-ouvertes.fr/hal-01100220

Un théorème sur l'équivalence de la valeur de H entre langages*

par Gabriel G. Bès

Les codes de réduction alphabétique proposent un système de codification des lettres d'un texte en langue naturelle qui tend à réduire le nombre de symboles binaires nécessaires à la représentation du texte, tout en sauvegardant la restitution de celui-ci de manière univoque. Les codes étudiés** sont fondés sur le principe de codifier non les symboles d'un alphabet unique, mais les symboles qui correspondent aux unités qui apparaissent dans les différents contextes et qui déterminent une pluralité d'alphabets: p. ex., on ne propose pas une codification de la lettre t, mais de la lettre t en position initiale de mot, en position intervocalique, en position initiale de syllabe précédée d'une consonne,etc. Dans chaque contexte, chaque unité sera associée à une séquence de symboles binaires tenant seulement compte des autres unités qui peuvent apparaître dans ce même contexte.

Certains des facteurs structuraux, très répandus dans les langues naturelles, qui déterminent les contextes où apparaissent les inventaires partiels des voyelles et des consonnes sont représentés, dans un texte, par des indices qui ne sont pas, nécessairement et toujours, organisés de gauche à droite. Soit, p. ex., dans une langue comme l'espagnol, la suite

$$C\ A\ \underbrace{M}_{x}\ \underbrace{\ldots}_{y}$$

Si dans la position y apparaît une voyelle, M dans la position x appartient à la catégorie des consonnes intervocaliques avec N, Ñ, P, R, S, T, etc.; en revanche, si dans la position y apparaît une consonne, M dans la position x appartient à un ensemble réduit de consonnes, les consonnes de fin de syllabe, d'où sont exclues beaucoup de celles qui figurent parmi les intervocaliques : les possibilités dans x sont conditionnées par les occurrences dans y; il est donc impossible de les prévoir à partir de la suite CA.

Or, si l'on souhaite profiter de cette situation, très largement attestée dans les langues naturelles, pour construire des codes de réduction alphabéti-

Condenser, n° 1, 1980, 97-100.

que, il est nécessaire d'obtenir, moyennant des règles, un "langage organisé" à partir d'un langage en langue naturelle (désormais, L = langage en langue naturelle, L' = langage organisé). Dans L', l'information (au sens non technique) sur la structure sous-jacente des énoncés, est véhiculée de gauche à droite, ce qui n'est pas toujours le cas dans L. Dans le cas de l'exemple, la suite CA... doit devenir la suite C'A'... de L'; à partir de la reconnaissance de A' de la suite C'A'... il faudra savoir si dans la position x suivante on aura une consonne intervocalique ou une consonne de fin de syllabe.

Pour obtenir L', on utilise un système de règles R opérant sur L. Parmi ces règles, il y en a qui font éclater un symbole de l'alphabet de L, et l'associent à deux ou plusieurs symboles de l'alphabet de L': un symbole x du langage L, p.ex., se réécrit X_1 lorsqu'il apparaît dans un contexte donné et X_2 lorsqu'il apparaît dans un autre contexte; on a:

$$X \longrightarrow X_1 \quad / \quad Z-$$
$$X \longrightarrow X_2 \quad / \quad Y-$$

C'est le cas de l'exemple précédent, où A doit se réécrire A_1 ou A_2 selon que dans la position y on a une voyelle ou une consonne.

D'autres règles de R vont associer une suite de deux ou plusieurs symboles de L à un seul symbole de L' (ou vice versa). Ce serait p. ex. le cas si dans une langue comme l'espagnol on introduisait la règle

$$q\frown u \longrightarrow q$$

le u étant la seule lettre possible à la suite de q. Grâce à ce type de règles, le nombre de symboles -lettres et/ou espaces- dans les suites de L peut être différent du nombre de symboles dans les suites correspondantes, associées par R, de L! Pourvu que les textes de L et de L' rapprochés par R aient une certaine longueur, il y aura une constante K permettant de trouver pour un texte quelconque, le nombre de symboles dans L' à partir du nombre de symboles dans L ou vice versa.

Il existe donc plusieurs types de règles R, qui se différencient selon les associations qu'elles déterminent entre unités de L et unités de L'. Dans tous les cas, cependant, à chacune des suites de L' doit correspondre une seule suite de L et vice versa. Entre les suites de L et celles de L', il existe donc une relation bijective.

Puisque c'est le langage L' qui doit être codifié, il se pose la question de savoir quel est le rapport de l'entropie de L avec l'entropie de L'. En particulier, il est clair que, par application des règles qui font éclater un sym-

bole de L en l'associant à deux ou plusieurs symboles de L', il y aura plus de symboles dans l'alphabet de L' que dans celui de L, ce qui risque d'entraîner l'augmentation de la valeur de l'entropie de L', et, par là, une codification nécessairement plus coûteuse de L' que de L, allant ainsi à l'encontre de l'objectif final poursuivi.

On sait que la valeur d'entropie d'un langage sera conditionnée par le type d'approximation selon lequel elle est mesurée : cette valeur variera selon que l'on considère seulement le nombre de symboles dans l'alphabet ou, en plus, leurs fréquences et, dans ce cas, selon que les fréquences correspondent à des symboles isolés ou bien à des symboles considérés dans les digrammes, trigrammes, n-grammes. On sait aussi que le nombre de symboles binaires utilisés pour codifier un texte est égal ou supérieur à l'entropie du langage multipliée par le nombre de symboles du texte. On doit donc se poser la question du rapport d'une certaine valeur d'entropie de L avec une certaine valeur d'entropie de L', et ceci en tenant compte du nombre de symboles qui apparaissent dans les suites de L et les suites correspondantes de L'. Plus précisément, dans le cadre de l'objectif final poursuivi, il importe de prouver que la transmission de L' n'est pas nécessairement plus coûteuse que la transmission de L; on désire donc prouver que

$$H_{\circ}\ N(1) = H'_{\circ}\ N(1')$$

où

H = entropie optimale de L

$H \leqslant H_i$ de L

H' = entropie optimale de L'

$H' \leqslant H_i$ de L'

N(1) = nombre donné de lettres et/ou espaces dans une suite finie de L

N(1')= nombre donné de lettres et/ou espaces dans une suite finie de L'

On sait:

(1) L' est obtenu de L par R; il existe une relation bijective entre L et L'.

(2) $K \cdot N(1) = N(1')$

(3) $G_N = -\frac{1}{N} \sum_i p(B_i) \log_2 (B_i)$

$$\lim_{N \to \infty} G_N = H$$

(Théorème prouvé par Shannon; G_N= entropie par symbole dans les suites de N symboles; cf. C.E.Shannon et W.Weaver, The mathematical theory of communication. Urbana, 1963, p. 24-25).

Soit: $\{B_1^L \ldots B_n^L\}$ l'ensemble de toutes les suites de N(1) dans L

$\{B_1^{L'} \ldots B_n^{L'}\}$ l'ensemble de toutes les suites de K.N(1) dans L'

$p(B_i^L) = p(B_i^{L'})$ (par (1))

$-\sum_i p(B_i^L) \log_2 p(B_i^L) = \sum_i p(B_i^{L'}) \log_2 p(B_i^{L'}) = X_N$

$\lim_{N \to \infty} \frac{X_N}{N(1)} = H$ (par (3))

$\lim_{N \to \infty} \frac{X_N}{K.\,N(1)} = H'$ (par (3))

$\frac{1}{K} \lim_{N \to \infty} \frac{X_N}{N(1)} = H'$

$\frac{1}{K} . H = H'$

$H.\,N(1) = K.\,H'.\,\frac{N(1')}{K}$ (par (2))

$H.\,N(1) = H'.\,N(1')$

ssi $K = 1$

$H = H'$

Le théorème précédent montre donc que si les règles R permettent d'obtenir un langage L' où à chacune des suites de L' correspond une seule suite de L et vice versa (cf. relation bijective entre L et L'), le coût de codification, mesuré en symboles binaires, des suites d'un texte quelconque, ne variera pas de L' à L si on réussit à proposer une codification de L et de L' dont les valeurs en symboles binaires soient respectivement identiques à H et à H'.

* Je remercie Michel Chambreuil de ses indications pour la mise au point finale de ce théorème.

** Cf. Gabriel G. Bès, "Structuration linguistique et mesure de l'information". In : Linguistique fonctionnelle - Débats actuels et perspectives. Hommage à André Martinet. Paris, P.U.F., 1979, p. 129-141.